Cognitive Behavior Therapy for Obsessive-Compulsive Behavior in Children with Autism Spectrum Disorder: A Comparison of Hierarchical Linear Modelling and Visual Analysis

Heather Yates, Masters of Arts

The Centre for Applied Disability Studies

Submitted in partial fulfillment

of the requirements for the degree of

Masters of Arts

Centre for Applied Disability Studies, Brock University

St.Catharines, Ontario

Abstract

Behavioral researchers commonly use single subject designs to evaluate the effects of a given treatment. Several different methods of data analysis are used, each with their own set of methodological strengths and limitations. Visual inspection is commonly used as a method of analyzing data which assesses the variability, level, and trend both within and between conditions (Cooper, Heron, & Heward, 2007). In an attempt to quantify treatment outcomes, researchers developed two methods for analysing data called Percentage of Non-overlapping Data Points (PND) and Percentage of Data Points Exceeding the Median (PEM). The purpose of the present study is to compare and contrast the use of Hierarchical Linear Modelling (HLM), PND and PEM in single subject research. The present study used 39 behaviours, across 17 participants to compare treatment outcomes of a group cognitive behavioural therapy program, using PND, PEM, and HLM on three response classes of Obsessive Compulsive Behaviour in children with Autism Spectrum Disorder. Findings suggest that PEM and HLM complement each other and both add invaluable information to the overall treatment results. Future research should consider using both PEM and HLM when analysing single subject designs, specifically grouped data with variability.

Acknowledgements

I would first and foremost like to thank Dr. Tricia Vause. You've done more for me than I could ever possibly express in one page. While working with you I developed my love for research. You made me answer the question '*why*?'. You taught me to not only think critically, but consider all of the options before making a decision. Thank you so much for including me in your decision making, and listening to my ideas – even when I was wrong. You're stuck with me now! I would also like to thank my co-supervisor, Dr. Jan Frijters for all of the work you put into the HLM analysis, as well as your time and patience as I asked you endless (and I do mean endless) questions about HLM. I would like to express my gratitude to my other committee member, Dr. Maurice Feldman, for your quick insightful feedback throughout this project. I would of course like to thank Naomi Johnson for all of the long hours you spent in the lab entering data, among other things. I am grateful to the parents and children who participated in this research project and their efforts in taking charge of OCB.

I feel so lucky to have been part of such an amazing program. The CADS program has so many strengths, one of which is Alison Rothwell. I think I can speak for my whole cohort in saying that we couldn't have made it through the program without our regular visits to your office. You truly are amazing!

This research would not have been possible without financial support from Brock University Graduate Studies, an Ontario Graduate Scholarship, and a New Investigator Fellowship from the Ontario Mental Health Foundation and the Ministry of Health and Long-Term Care.

Table of Contents

List of Tables

Cognitive Behavior Therapy for Obsessive-Compulsive Behavior in Children with Autism Spectrum Disorder: A Comparison of Hierarchical Linear Modelling and Visual Analysis

Introduction

Single subject research designs are commonly used in behavior analytic research (Cooper, Heron, & Heward, 2007; Kazdin, 2011; Rodriguez, Thompson, Schlichenmeyer, & Stocco, 2012). Methods for analysing single subject research have continuously evolved in an attempt to quantify research results and limit intrinsic methodological drawbacks (Davis et al., 2013). Increased knowledge in effective treatment practices has translated to more complex research designs, with larger groups of participants (e.g., RCT's) (Davis, 2012). To accommodate these changes, more sophisticated statistical techniques are needed. Specifically, a technique is needed that can evaluate the statistical certainty of effect calculations (Parker & Brossart, 2006). For the sake of simplicity, the present study will use the term visual analysis to talk about the percentage of non-overlapping data points (PND) and the percentage of data points exceeding the median (PEM), and the term visual inspection when talking about visual inspection. This is important to note because visual inspection can technically be classified as a form of visual analysis (Davis et al., 2013).

Percentage of Non-overlapping Data Points

Current analyses in behavioural research have included: (a) visual inspection; and (b) visual analysis (Percentage of non-overlapping data points [PND]); Percentage of data points exceeding the median [PEM]) (Ma, 2006; Preston and Carter, 2009). Visual inspection, which is commonly used by behaviour analysts, is a method of analyzing data which assesses the variability, level, and trend both within and between conditions. Using single case experimental design with visual inspection has good internal validity, and is able to show experimental control

(Cooper et al., 2007). PND was formulated as a means of quantifying results in a way that visual inspection is unable to do (Parker, Hagan-Burke, & Vannest, 2007). PND is the most frequently used form of visual analysis (Bellini & Akullian, 2007; Lee, Simpson, & Shogren, 2007; Preston & Carter, 2009). It is calculated by determining the percentage of data points in the treatment phase, which are below (or above when the treatment goal is to increase a behaviour or skill) the lowest (or highest) baseline data point (Preston & Carter, 2009). The percentage of non-overlapping data points is then evaluated on a scale which determines treatment effectiveness. The scale is as follows: *91% - 100%* (highly effective), *71%-90%* (moderately effective), *51%-70%* (questionably effective), and *50% and below* (ineffective) (Chen & Ma, 2007; Ma, 2006; Scruggs & Mastropieri, 1998).

In addition to a standard evaluation scale (Chen & Ma, 2007), PND is easy to calculate, understand and interpret (Browder, Spooner, Ahlgrim-Delzell, Harris, & Wakeman, 2008; Ma, 2006). One reason why PND is often chosen over other inferential statistics is because it does not require the assumption of independence, whereas most statistical techniques do (e.g., *t*-test) (Bellini & Akullian, 2007; Ma, 2006). This is important given that many single subject designs violate this assumption resulting in a statistic that is nonparametric and does not have known distributional characteristics or associated probabilistic test-statistics (Bellini & Akullian, 2007).

Due to the popularity of PND, several meta-analyses have been conducted using PND to analyze data (e.g., Bellini & Akullian, 2007; Browder & Xin, 1998; Didden, Korzilius, VanOorsouw, & Sturmey, 2006; Marthur, Kavale, Quinn, Fornuss, & Rutherford, 1998). Bellini and Akullian (2007) conducted a meta-analysis of 55 single subject design studies published between 1986 and 2005, which evaluated a school-based intervention for children and adolescents (N = 157) with Autism Spectrum Disorder (ASD). PND was calculated for all

studies and the researchers found a questionable intervention effect (PND *M=70%,* range*=17-100%).* Though PND is widely used, it is also criticized (Chen & Ma, 2007; Davis et al., 2013; Ma, 2006). One common critique of PND is that even if one baseline data point reaches the ceiling or floor (depending on the direction of the data), the PND score will be 0% meaning no treatment effect (Ma, 2006). Therefore, one data point can determine whether a treatment considered is effective, or not. This is problematic, particularly in clinical data sets, where data is notoriously variable (Mitchell, 2012). In addition, PND scores are often highly correlated with the number of data points; the more data points present, the higher the likelihood the treatment presents as effective (Davis et al., 2013).

Percentage of Data Points Exceeding the Median

The acknowledged challenges with PND left a need for a new way to quantitatively analyze data. The need resulted in a technique called PEM. PEM determines treatment effectiveness by calculating the percentage of treatment data points which exceed the median baseline data points (Ma, 2006). PEM uses the same easy to understand scale as PND, which is easy to calculate, and interpret, without violating the assumption of independence (Chen & Ma, 2007; Ma 2006). PEM on the other hand is not affected by outlying data points like PND as it uses the median baseline score instead of the lowest score. Authors have argued that PEM is a good alternative to address some of the inherent weaknesses seen in PND (Chen & Ma, 2007; Ma, 2006).

Though PEM is a relatively new form of data analysis, several meta-analyses have used PEM (e.g., Chen & Ma, 2007; Gao & Ma, 2006; Ma, 2006; Ma, 2009). A meta-analysis by Ma (2006) comparing PND and PEM was completed. He looked at the data in 16 articles (e.g., Feldman, Ducharme, & Case, 1999; Levendoski & Cartledge, 2000; O`Reilly, Green, & Braunling-McMorrow, 1990), including 659 pairs of baseline and treatment phases, using visual

inspection of single subject research in self-control (Ma, 2006). This author applied both PND and PEM to the 659 graphs, finding that PEM with a score of 0.87, (*SD*=0.24) had a higher correlation with the original author's judgements, compared to PND which had a score of 0.61, (*SD*=0.39) (Ma, 2006). Within this study, PND showed that the intervention was questionably effective, whereas PEM suggested the intervention was moderately effective. This is an important distinction, given that studies often code questionably effective and ineffective together (Chen & Ma, 2007; Gao & Ma, 2006; Ma, 2006). These findings suggest that PEM may be important, specifically in uncovering a treatment effect that PND may have missed due to using the lowest data (or highest if learning a skill) point in baseline.

Though PEM has had relatively limited application in research, it has shown promise in addressing some of the limitations seen in PND (Chen & Ma, 2007; Ma, 2006). This being said, several limitations still exist. One critique with PEM is that it is not sensitive to magnitude or clinical significance. For example, behaviour on a 5-point scale may have mostly scores of 5, with a few 4's, in baseline, and then all 4's in the treatment phase. The PEM score, for a situation like this, would be 100%, meaning that it is a highly effective intervention (Ma, 2006). Other limitations to PEM is that similar to PND, PEM is highly correlated with the number of data points in the treatment phase and they do not account for tend in the baseline phase. Overall, some researchers have found that visual analysis can be imprecise (Brossart, Parker, Olsen, & Mahadevan, 2006), and may only catch the stronger treatment effects, missing some of the smaller, equally important ones (Brossart et al., 2006; Davis et al., 2013; Kromrey & Foster-Johnson, 1996). For this reason, Davis et al. (2013) suggest that Hierarchical Linear Modelling (HLM) be considered to assist in quantifying data.

Hierarchical Linear Modelling in Single Subject Design

HLM is a multilevel statistical model which can analyze multiple levels of data, over time, between and within phases (Field, 2013). For example, on level 1, you can investigate overall effectiveness of a treatment package (within participants and across phases), and on a superordinate level such as level 2 (e.g., between participants), you can compare the impact that predictor variables (e.g., age, IQ, or response class) have on the sample's treatment outcomes. In order to use HLM, three criteria must be met. The data must have at least 3 waves (e.g., three assessments), an outcome variable is needed in which the values change in a systematic way, over time, and a reasonable and reliable method of time is used (e.g., days, months, treatment occasions, etc.) (Singer & Willet, 2003). HLM includes three important components: intercept, slope, and variance. An HLM estimates the intercept, which represents the average outcome across all participants, depending on how the repeated measures are parameterized (e.g., linear/curvilinear growth over time; piecewise growth parameters; a step function representing global differences between phases in a single-subject design). HLM also gives several estimates of person to person variability, estimating both meaningful variability in intercept, slope and their covariance, along with an estimate of residual variability not accounted for by the model. Critical to the current application, HLM models can generate an estimate of treatment effect for each person, along with an evaluation of whether treatment effects systematically vary by some individual difference variable (e.g., age, response class, etc). Overall, HLM provides a rich picture of therapeutic outcomes that, among other things, can assess overall treatment as well as individual treatment effects and the person-level factors that predict variability in those effects.

Past single subject studies have largely stayed away from traditional inferential statistics (e.g., Analysis of Variance (ANOVA) and *t*-tests) partially due to the time series research

designs leading to a violation of many basic statistical assumptions. Three of these assumptions that are often problematic include: independence of observations, the need for random selection of participants, and normal distribution with constant variance (Davis et al., 2013). HLM, however, explicitly models the dependence among the multiple observations for a given participant, adjusting the standard errors of any test of treatment effect accordingly. Even more interesting, unlike *t*-tests, where mean differences are compared, HLM allows for change across time to be taken into account via formulation of a model for change that is informed by study design and therapeutic or developmental patterns of change (Singer & Willet, 2003). Lastly, unlike PEM and PND which may confound the number of data points with the size of the effect, the HLM model uses the number and variability of within and across phase data to increase the precision of mean estimates for each phase.

HLM research has been extended to include analysis in single subject research (e.g., Lumpkin, Silverman, Weems, Markham, & Kurtines, 2002). One study used group cognitive behavioural therapy (CBT) to treat anxiety in 12 typically-developing children and adolescents (aged 6-16 years). The authors found that when looking at parent ratings of daily severity, there was no significant difference in the between-subject variables used in level 2, but did find a significant linear trend between the end of the baseline phase and the end of the intervention phase, $X2 = 67.61$, $p < .01$, (Lumplins et al., 2002). The authors argued that HLM was a good fit for similar data sets because it not only allows for multiple levels of analyses, but also allows for variable amounts of time between observations (Lumpkin et al., 2002). Though HLM has much strength, it has some shortfalls as well. HLM can be very challenging to both calculate and interpret. In fact, Kratochwill et al. (2010) argue that multilevel models are the least understood of all analytic procedures due to the complexity of the analyses. This can be an issue for both the

researcher and the reader. HLM also does not calculate an effect size, and needs more data points, for statistical power, compared to visual analysis. Moreover, limited research exists to validate the use of HLM in single subject designs.

HLM was recently used to complement visual inspection in a study, which investigated the effect of a literacy intervention on student's sight word acquisition, in a multiple baseline design across participants with an embedded changing criterion design (Davis et al., 2013). Participants included 11 students with an intellectual disability. The researchers found that both visual inspection and HLM added invaluable information to the overall results. Visual inspection was able to determine a functional relationship between the introduction of the literacy program and the acquisition of sight words, and HLM was able to add statistically significant to this relationship as well as identify variables (e.g., print knowledge) which helped to predict student success. Overall, the authors suggest using both visual inspection and HLM when analyzing data, in order to gain the benefits of both techniques.

To date, only one study has compared HLM to visual inspection (Davis et al., 2013), and no studies have compared HLM to visual analysis. The present study will expand on this limited research and compare treatment outcomes using PND, PEM, and HLM on three response classes of Obsessive Compulsive Behaviour in children with Autism Spectrum Disorder (ASD).

Obsessive Compulsive Behaviours in Children with Autism Spectrum Disorder

ASD is a neurological disorder which is characterized in the Diagnostic and Statistical Manual-V-TR (*DSM-V-TR*) by symptoms in three key domains: qualitative impairments in communication, social interaction, and the presence of repetitive and/or restricted behaviors (American Psychiatric Association [APA], 2013). The prevalence of ASD, among children, has consistently risen over the past several decades. In 2006, the prevalence of ASD in children and

youth was estimated to be 1 in155 in Canada (Fombonne, Zakarian, Bennett, Meng, & McLean-Heywood, 2006) and approximately 1 in 110 in the US (Centers for Disease Control and Prevention, 2006). More recently, this estimate has risen to 1 in 50 in the US (Centers for Disease Control and Prevention, 2013).

Researchers (e.g., Hollander et al., 2009) have conceptualized repetitive behaviours as falling into two distinct categories: higher-order repetitive behaviours and lower-order repetitive behaviours. Behaviors in the higher-order category may serve to reduce anxiety; whereas, lower-order behaviours (e.g., handflapping, self-injury) are often perceived to moderate arousal (Hollander et al., 2009). Higher-order behaviours including ordering, washing, checking behaviours, and rituals (e.g., regimented bedtime routine) often overlap with behaviors characteristic of Obsessive-Compulsive Disorder (OCD; APA, 2013). In a study by Miranda et al. (2010) that analyzed the Repetitive Behavior Scale-Revised (RBS-R) (Bodfish, Symons, & Lewis, 1999) compulsive, ritualistic, and sameness subscales were all held within the same factor. This suggests that these behaviors are similar to each other and different from other categories of repetitive behaviour (e.g., self-injurious behavior). Distinguishing between higher-order behaviours that may be anxiety-driven versus those controlled by other functions (e.g., automatic positive reinforcement) can be challenging given the social and communicative deficits that are often seen within ASD (Reaven et al., 2009). The presence of obsessions are often challenging to uncover either due to lack of insight into thoughts, and/or inability to verbalize them (Gillott, Furniss, & Walter, 2001). Given this challenge, the present study will use the term Obsessive-Compulsive Behaviour (OCB).

To date, limited research has focussed on exclusively treating obsessive-compulsive behaviour in children with ASD. Randomized Control Trials (RCTs) have been completed to

evaluate CBT to treat anxiety (including OCD) in children with ASD (e.g., Storch et al., 2013; Wood et al., 2009), but have included few participants with OCD and individual treatment outcomes are not disclosed. In addition, three uncontrolled case studies (N = 1) have been evaluated adapted CBT packages to address OCD in children with ASD (Lehmkuhl, Storch, Bodfish, & Geffkin, 2008; Reaven & Hepburn, 2003; Sze & Wood, 2007) where participants showed treatment remittance. Last, behavior analytic treatment has been evaluated with single subject designs and visual inspection (e.g., Rodriguez, Thompson, Schlichenmeyer and Stocco, 2012; Sigafoos, Green, Payne, O'Reilly, & Lancioni, 2009), and has shown clinically significant decreases in OCBs.

An ongoing RCT (Vause et al., 2013) involves a 9-week manualized group CBT treatment package titled, "*I Believe in ME, Not OCB!,"*(Vause, Neil, Yates, & Feldman, 2013a) to treat obsessive-compulsive behaviours in children with a diagnosis of ASD. This treatment adapts traditional CBT to meet the needs of a pediatric ASD population (e.g., increased parent involvement, use of visuals). A preliminary study completed by Neil (2011) piloted this package to treat OCBs in four children (ages 7-11 years) with ASD and OCB. A multiple baseline design across behaviours was used with daily parent ratings (Cooper et al., 2007). Visual inspection indicated a clinically significant drop in parent ratings of OCBs when the active treatment was introduced. Clinical significance was achieved when parent goals for behaviour outcomes were met.

The present study will extend this research by analyzing results according to three response classes with three separate forms of data analysis (PND, PEM, HLM). Three methods are being used because previous research has focused on PND, but more recent researchers have begun to note advantages to using PEM. Several studies have compared the PEM and PND,

concluding that PEM addresses limitations found of PND (Ma, 2006). HLM is being used in an attempt to explore the benefits of using a statistical tool in single subject designs. This is the first known study to compare all three methods.

Hypotheses

My **hypotheses** are:

1. Hierarchical Linear Modeling will detect more within-subject treatment effects, compared to PEM and PND.
2. PEM will detect more within-subject treatment effects, compared to PND.
3. There will be moderate, but varying correlations among the three methods for quantifying treatment effects.

Method

Participants

In total, 35 children (18 experimental and 17 control) who were 7 to 13 years of age, and at least one parent voluntarily participated in the RCT. In the present study, a total of 39 behaviours (1-3 behaviours per child) were analyzed across 18 participants selected from the RCT. Inclusion criteria included an ASD diagnosis (APA, 2013), and the presence of obsessive-compulsive behaviours according to the Children's Yale-Brown Obsessive Compulsive Scale (CY-BOCS; Albano & Silverman, 1999) and RBS-R (Bodfish, Symons, & Lewis, 1999). In addition, children needed to be between 7-12 years of age at intake, and have an estimated IQ score of >70 according to the Wechsler Intelligence Scale for Children-Fourth Edition (Weschler, 2004). In the present study, the children's IQ's ranged from (60-110)

Response Classes

The present study examined three response classes of target behaviours. Response classes included: contamination-related behaviours, same seat behaviours, and arranging and ordering

behaviours. These response classes were chosen in two ways. The first criterion for choosing response classes was that a minimum of 10 behaviours were needed. This criterion was used in order for there to be enough power to complete HLM analyses. The three chosen response classes had the greatest number of behaviours in the RCT. There were 14 contamination related behaviours, 14 same seat behaviours, and 11 arranging and ordering behaviours. See Table 1 for a listing of behaviours from each of the three OCB response classes. Looking specifically at each class, all compulsions in the contamination class had an obsession connected to them (or others) related to becoming contaminated (e.g., dirty, sick). Arranging and ordering behaviours had the same or similar topography; the compulsion involved the child needing to arrange items, and not allowing others to change their arrangement. Last, behaviours in the same seat response class were similar in topography. In each case, the child needed to sit in a particular seat (e.g., in the car, on the couch, at the kitchen table, etc). Through parent rating questionnaires, data was collected daily. Questionnaires included a minimum of one question (more if deemed necessary) for each target behavior. Parents answered each question by checking a box on a Likert scale. Across all behaviours, in all classes, a score of 5 represented the behaviour at a *undesirable* (interfering) level, where the parent observed it to be at the beginning of treatment, a score of 1 represented the *desirable* level (parent's goal), and a score of 3 represented approximately halfway in between the undesirable level at the start of treatment and the parents goal. For example, if the child is currently rinsing their mouth 8 times every morning and the parents goal is for the child to rinse their mouth only 1 time than the child rinsing their mouth 4 or 5 times is about half way between the two extremes. When deemed necessary (e.g., the frequency of the behaviour was appropriate to collect data on) 5 anchors were used, though the basic framework

for anchors remained consistent. See Table 2 for examples of anchors used with behaviors in the three response classes.

Table 1

Sample Target Behaviours

Response Class	Sample Behaviours
Contamination response class	Handwashing
	Making comments about food being rotten, or having germs
	Rinsing mouth several times after brushing teeth
Same seat response class	Sitting in the same seat at the kitchen counter
	Sitting in the seat behind the passenger in the car
	Sitting in the same seat at the kitchen island
Arranging and ordering response class	Arranging items (e.g. LEGO, Mario, Stuffed animals) on the shelves
	Arranging (e.g. stickers, butterflies) on the computer table
	Arranging stuffed animals in a straight line

Table 2

Measurement of Target Behaviours

Sample behaviours	**1** **Desirable level (parents goal). Behaviour is not interfering**	**2**	**3** **(Halfway between 1 and 5)**	**4**	**5** **Undesirable level (current level at beginning of treatment) Behaviour is interferring**
After brushing his teeth in the morning, how many times did X need to rinse his mouth?	1 or less times	2-3 times	4-5 times	6-7 times	8 or more times
While in the car, was X able to sit in a seat other than directly behind the passenger seat?	Yes, he was		Some of the time		No, he was not
If items were moved in X's dinosaur collection, how well did he tolerate the movements?	Tolerated several movements		Tolerated different positions of one item		Has to move it back

Treatment and Materials

The present study was approved by the Research Ethics Board, and informed consent and assent forms were signed. The treatment was a manualized group function-based cognitive behavioural therapy package titled "*I Believe in Me, Not OCB!"*. The package included a children's workbook (Vause, Neil, Yates, & Feldman, 2013a) and a clinician's manual (Vause, Neil, Yates, & Feldman, 2013b). The intervention was nine sessions, each lasting 2 hours. Each group included two to four children, at least one parent for each child, and two therapists. The key treatment components are: (a) Psychoeducation and Mapping, and (b) Functional Behavioural Assessment Intervention (FBA/I), (c) Cognitive Behavioural Therapy (CBT) skills training, (d) Exposure and Response Prevention (ERP). Psychoeducation and mapping (PM) is accomplished though various means including, mapping behaviors into one of three zones, using a thermometer scale to rate severity of behaivours (1-5), drawing pictures, probing exercises, and group discussion. Moreover, this phase introduces the families to terminology that will be needed throughout treatment. The FBA/I component utilizes the Questions about Behavioural Function (QABF) assessment (Matson & Vollmer, 1995) to uncover possible functions which may maintain the target behaviour. Strategies for how to address each function are given to parents each week. The CBT skills training component involves teaching children coping statements (e.g., externalizing statements and positive self-statements, cognitive restructuring, narrative roleplaying). Last, ERP consisted of gradual exposure to feared stimuli through planned exposures and loose blocking of the compulsion. In addition to these key treatment components, there was also a weekly parent training session, and a social skills component. The parent training session was 30 minutes each session, and included topics such as schedules of reinforcement, cognitive therapy, exposure and response prevention, and role playing. The social

skills component involved social games (e.g., cheeky chance, alphabet actions). For a thorough description of the treatment package, refer to Neil (2011, unpublished).

Reliability

Inter-observer agreement (IOA) of PEM and PND analyses were calculated by dividing the number of agreements by the number of agreements plus the number of disagreements. IOA was calculated randomly, by a naïve observer, on 10 of the 39 behaviours used in the present study. Both PEM and PND had 100% agreement between the original researcher and the individual who calculated IOA.

Treatment Integrity

Treatment integrity (TI) for treatment sessions was calculated for each rating as the number of agreements dividing the number of agreements by the number of agreements plus the number of disagreements, multiplied by 100% (Cooper et al., 2007). All treatment sessions were video recorded. TI was completed by using a treatment checklist on a randomly selected 55% of sessions. TI was scored by several scorers. The mean treatment integrity was 100%. Treatment Integrity checks were completed on 18% of session videos with 100% agreement. Reliability was calculated as the number of agreements divided by the number of agreements plus disagreements multiplied by 100% (Cooper et al., 2007).

Analyses

Three types of analyses were completed on this data set: HLM, PND, and PEM. Visual inspection was done and found that the data was frequently variable. Given the variability in the present study`s data set, it's likely more suited to PEM. The author has chosen to include PND as well due to the long history of using this technique throughout the literature. Also, in the results section, the term *"person-behaviour pair"* will be used to represent the combination of a person

and his/her behaviour. This is needed because some participants have more than one behaviour and each person-behaviour pair is analyzed separately. For all analyses, the study was separated into two phases. The first phase is the baseline-PM phase. This phase includes parent ratings during baseline and PM. The second phase (or what was deemed the "active treatment" phase) includes the FBAI, CT, and ER/P components. When visually inspecting the baseline and PM data, the author saw minimal differences in the data. Further, studies (e.g., Vause, Neil, Yates, Jackiewicz & Feldman, in progress) did not show a statistically significant difference between baseline and PM. In HLM, the two different levels of analyses were run: unconditional growth model and conditional growth model. The unconditional growth model is the most basic model and does not include any covariates. This model asks the question "what are the differences between the before treatment was introduced and after treatment was introduced?". In the conditional growth model a covariate was added (response class). This model asks the question "why does this difference exist?". To incorporate response class as an explanatory predictor, the three response classes were dummy coded into two vectors that could predict variability in baseline-PM levels and step function. Arranging and ordering was arbitrarily chosen as a reference category and dummy codes were assigned such that dummy vector 1 (d1, below) represented the deviation of contamination from arranging and ordering, while dummy vector 2 (d2, below) represented the deviation of same-seat behaviours from arranging and ordering.

The graphs used to complete PEM and PND analyses were originally in a multiple baseline design across behaviours. Graphs were examined by the author and showed experimental control. When defining whether a treatment was effective, or not, according to PEM or PND, the behaviours were first organized into one of four categories: *91% - 100%* (highly effective), *71%-90%* (moderately effective), *51%-70%* (questionably effective), and *50%*

and below (ineffective) (Chen & Ma, 2007; Ma, 2006; Scruggs & Mastropieri, 1998). A precedent has been set in previous research (Chen & Ma, 2007; Gao & Ma, 2006; Ma, 2006), when comparing methods of analysis, to consider *questionably effective* and *ineffective* as both being not effective. Therefore, in the present study, 70% and below were scored as being ineffective and 71% and above were scored as effective.

Results

Visual Analysis

Contamination response class.

The contamination response class included 14 behaviours, across 7 participants. All scores for each of the three methods of analysis are reported in Table 3. According to an analysis completed using HLM, there was a statistically significant decrease between the baseline-PM phase and the treatment phase for all 14 behaviours. Analyzing the data with PEM resulted in a range of scores between 0.28 (*ineffective*) and 1(*highly effective*). Treatment was effective in reducing the target behaviour in 13 of the 14 behaviours. Behaviour 14 was the only behaviour where PEM does not show a treatment effect. Upon visually examining the graphs, this appears to be due to limited data in the treatment phase as the baseline-PM phase is stable at a score of 5, and the treatment phase is variable between a score of 1 and 5. Last, PND indicated a treatment effect for 7 of the 14 behaviours treated in the contamination response class. The PND score ranged from 0 (*ineffective*) to 0.96 (*highly effective*). After visually examining the graphs, it appears that 5 of the 7 behaviours which PND deemed treatment as not being effective for were cases where a small number (often only 1) of baseline-PM points dropped to a 1 or 2.

Table 3

Contamination Response Class Analysis Table

Behaviours	PND	PEM	HLM
1	0.81**	0.81**	-3.63*
2	0.92**	0.92**	-3.46*
3	0	1**	-0.77*
4	0	1**	-0.6*
5	0.35	0.85**	-1.26*
6	0.36	0.85**	-1.54*
7	0.96**	0.88**	-2.51*
8	0.47	0.81**	-1.55*
9	0	1**	-0.8*
10	0.92**	0.92**	-3.21*
11	0.76**	1**	-2.71*
12	0.74**	1**	-2.7*
13	0.73**	0.92**	-2.27*
14	0.55	0.55	-3.27*

*$p<0.01$, one-tailed. . **Effective treatment response using visual analysis.

Same seat response class.

The same seat response class included 14 behaviours across 8 participants. All scores for each of the three methods of analysis are reported in Table 4. According to an analysis completed using HLM, there was a statistically significant decrease between the baseline-PM phase and the treatment phase, for 13 of the 14 behaviours treated. Analyzing the data with PEM resulted in a range of scores between 0.16 (*ineffective)* and 0.96 (*highly effective).* The treatment package was effective in reducing the target behaviours in 10 of the 14 behaviours. PEM did not show a treatment effect for behaviours 8, 10, 11, and 12. Upon visually examining the graphs, this appears to be due to two key factors: limited data in the treatment phase or significant variability in the treatment phase. For behaviour 12, HLM and PEM are in agreement that no treatment effect occurred. Last, PND indicated a treatment effect for 5 of the 14 behaviours treated in the same seat response class. The PND scores ranged from 0 (*ineffective)* to 0.95 (*highly effective).* After visually examining the graphs, it appears that seven of the nine

behaviours which PND deemed treatment as not being effective resulted in a PND score of 0 due to a baseline-PM point dropping to a score of 1.

Table 4

Same Seat Response Class Analysis Table

Behaviours	PND	PEM	HLM
1	0.8**	0.8**	-3.81*
2	0.79**	0.79**	-3.33*
3	0.86**	0.86**	-3.29*
4	0	0.8**	-2.66*
5	0.82**	0.86**	-1.87*
6	0	0.79**	-1.69*
7	0	0.95**	-2.42*
8	0.22	0.28	-1.12*
9	0.95**	0.95**	-3.69*
10	0.44	0.44	-1.76*
11	0	0.3	-1.18*
12	0	0.16	-0.38
13	0	0.96**	-1.79*
14	0	0.93**	-3.25*

*$p<0.01$, one-tailed. **Effective treatment response using visual analysis.

Arranging and ordering response class.

The arranging and ordering response class included 11 behaviours across 7 participants. All scores for each of the three methods of analysis are reported in Table 5. According to an analysis completed using HLM, there was a statistically significant decrease between the baseline-PM phase and the treatment phase for 6 of the 11 behaviours. Analyzing the data with PEM resulted in a range of scores between 0.28 (*ineffective*) and 1 (*highly effective*). Treatment was effective in reducing the target behaviour in 5 of the 11 behaviours. HLM and PEM are in agreement with 5 of the 5 behaviours where a treatment effect was not found. Lastly, PND indicated a treatment effect for 4 of the 11 behaviours treated in the arranging and ordering response class. The PND score ranged from 0 (*ineffective)* to 1 (*highly effective)*. After visually examining the graphs, it appears that PND is consistent with HLM in 5 of the 7 cases where a

treatment effect was not found. In the remaining 2 cases, this appeared to be due to a baseline-PM point dropping to a score of 1.

Table 5

Arranging and Ordering Response Class Analysis Table

Behaviours	PND	PEM	HLM
1	0.88**	0.88**	-1.17*
2	0	0	-0.53
3	0	0	-0.34
4	0	0.93**	-1.09*
5	1**	1**	-3.55*
6	0	0.53	-2.5*
7	0.77**	0.77**	-0.7*
8	0	0	0.82
9	0	0	0.02
10	0	0	0.17
11	0.77**	0.77**	-1.95*

*$p<0.01$, one-tailed. **Effective treatment response using visual analysis.

Response class comparison summary.

In summary, of the 39 behaviours utilized in this study, the treatment was effective for 15 according to PND, 28 according to PEM, and 33 according to HLM (See Tables 3, 4, and 5).

Hierarchical Linear Modelling

The following section outlines the results found when analyzing the data set first with an unconditional growth model and then second with a conditional growth model. Note that the conditional growth model is the model where the response class predictor was introduced to the HLM analyses.

Unconditional growth model.

The level 1 model used in the following unconditional growth model is:

$$Yij = [\beta_{00} + \beta 10\ STEPij] + [uoj + u1jSTEPij + rij]$$

The average baseline-PM severity score across all person-behaviour pairs was significantly greater than zero, β_{00}=3.97 (*SE*=19), $t(38) = 21.07$, $p < .0001$. There was also

statistically significant variability in baseline-PM scores among person-behaviour pairs $\boldsymbol{\tau}_{00}$=1.35, (*SE*=0.32), $z = 4.25$, $p < .0001$. This means that although there is a statistically significant difference from zero (the lowest possible score) for the average person-behaviour pair score, there was also meaningful variability among person-behaviour pairs during the baseline-PM phase. An explanation of what accounted for this variance was further investigated during the level 2 analysis.

The baseline-treatment difference in severity scores across all person-behaviour pairs was statistically significant $\beta_{10} = -1.88$ (*SE*=0.21), $t(2933)=-9.14$, $p< .0001$. This result indicated that across all person-behavior pairs each behaviour improved by 1.88 points, accounting for the repeated measurements over time. In addition to this, the person-behaviour variability in step function was statistically significant $\boldsymbol{\tau}11=1.59$ (*SE*=0.39), $z=4.13$, $p< .0001$, meaning that there was meaningful and statistically significant variability across person-behaviours in the size of the behavioural improvement. The analysis also revealed that intercept variance was negatively correlated with variability in step function, $\boldsymbol{\tau}_{10} = -0.81$ (*SE*=0.28), $z = -2.87$, $p = .0042$. This indicated that the higher the person-behaviour pair began in baseline-PM, the greater the improvement they made in decreasing target behaviours. After all variability was accounted for, the remaining residual was $\sigma^2 = 0.79$ (*SE*=0.02), $z=38.03$, $p < .0001$. This indicated that meaningful person-to-person variability remained after accounting for the fixed effects (i.e., β_{00}, β_{10}), suggesting that additional fixed effects predictors could be added to the model.

Conditional growth model

The model used in the following conditional growth model is:

$$Yij = \beta_{00} + \beta_{10}\,(STEP)ij + \beta_{01}\,(RCj - RC) + \beta_{11}(RCj - RC)(STEP)ij + uoj + u1j\,(STEP)ij + rij$$

Adding the response class covariate to the HLM model allows us to investigate whether variation in the slopes and intercepts is related to which response class a behaviour is in. When response class was added, variability in baseline-PM scores among person-behaviour pairs remained statistically significant $\boldsymbol{\tau}_{00}$=1.20, (*SE*=0.29), $z = 4.14$, $p < .0001$. This means that there was less variability between person-behaviour pairs during the baseline-PM phase in response classes than between response classes. In addition to this, the person-behaviour variability in step function decreased, but remained statistically significant $\boldsymbol{\tau}11$=1.32 (*SE*=0.33), z=4.00, $p< .0001$, indicating that there was meaningful and statistically significant variability across person-behaviours in the size of the behavioural improvement, but that this variability was slightly less within response classes. This decrease from $\boldsymbol{\tau}11$=1.59 to $\boldsymbol{\tau}11$=1.32 means that response class accounts for 27% of the variation in target behaviours.

When comparing response classes, the analysis showed how the behaviours in each response class differentially responded to treatment. Recall that arranging/ordering was chosen as a baseline-PM response class, with two dummy-coded vectors coded to represent deviation from this response class by Contamination (d1) and Same Seat (d2). The treatment-baseline difference in severity scores across the arranging and ordering response class person-behaviour pairs was statistically significant β_{10}(RCd0) = -0.94 (*SE*=0.36), $t(2931)= -2.62$, $p < .0089$. This result indicated that on average each behaviour improved by 0.94 points. The baseline-treatment difference in severity scores between contamination versus arranging/ordering was also statistically significant, β_{10}(RCd1) = -1.23 (*SE*=0.48), $t(2931)= -2.59$, $p < .0096$ indicating that its average behaviour decreased by 1.23 more than arranging and ordering, or 2.17 points. Lastly, the baseline-treatment difference in severity scores for the same seat versus arranging/ordering was also statistically significant, β_{10}(RCd2) = -1.38 (*SE*=0.48), $t(2931)= -2.90$, $p< .0037$

indicating that its average behaviour decreased by 1.39 more than arranging and ordering, or 2.33 points. Post-hoc analyses indicated no differences between contamination and same seat behaviours, either in intercept or baseline-treatment severity.

Correlations

Pearson correlations were done to compare the treatment outcomes of all three data analysis tools. The correlation between PEM and PND is $r = 0.42$. When comparing HLM with PND, the correlation is $r = 0.69$. Lastly, when comparing PEM and HLM, the correlation was $r = 0.65$.

Discussion

When discussing outcomes from the three methods of analysis used in the present study, particularly treatment outcomes, one important factor to consider is choosing the most applicable method of analysis. Results of analyses for treatment outcomes are used to empirically validate effective treatments for use in applied clinical settings. Therefore, accurate results that provide the most possible information are of the utmost importance. The current study investigated whether there was an increase or decrease in the severity of target behaviours, in three OCB response classes, with the onset of active treatment (the introduction of FBA/I, CT, and ERP). The present study examined whether the treatment was effective for each person-behaviour pair as well as overall response classes. Overall, results indicate that the treatment package was effective in decreasing OCBs in the current sample of children. This finding is in line with previous research (e.g., Storch et al., 2013; Wood et al., 2009).

Overall there were 39 behaviours. HLM did not find a treatment effect for 6 behaviours, PEM did not find a treatment effect for 11 behaviors and PND did not find a treatment effect for 24 behaviours. When comparing PND, PEM, and HLM, one important finding is that the number

of behaviours where PND and HLM found that the treatment was not effective for was very different. This large discrepancy can mean the difference between deciding whether a treatment package is effective or ineffective. In 17 of the 24 cases where PND found no treatment effect, the PND value was 0%. In almost all of these cases, the zero was a result of one baseline-PM data point dropping to 1 (the lowest possible score). PND calculates the percentage of data points below the lowest point, so when a baseline-PM point drops to 1 (which frequently happens in clinical data sets), PND will not detect even the strongest of treatment effects. Parent ratings were also restricted to a 5-point scale, which sometimes only included 3 options (e.g., Yes, sometimes, no). This may have resulted in more scores of 1 than a larger scale would have. For these reasons, as expected, PND does not appear to be well suited for this data set.

PEM and HLM not only individually made unique contributions to the analysis of this data set, but they were also used to support each other's findings. Visual analysis and HLM both concluded that at the beginning of treatment, the children had behaviours at a clinically significant level that was statistically different from zero. Overall, PEM found a clear drop in the behaviours between the end of the baseline-PM phase and beginning of the treatment phase. HLM supports this finding and added an estimated value to this drop, 1.88. This means that on average, across all person-behaviour pairs, behaviours dropped by 1.88 points on the parent rating scale. HLM also revealed that there was statistically significant variability in the amount of decrease between baseline-PM and active treatment. This indicates that the magnitude of the change was varied systematically across person-behaviour pairs. This is an important contribution because one of the limitations of PEM is that it does not account for the magnitude of change.

An important contribution that HLM adds to single subject research is that it provides a way of explaining the variability in treatment effects, something that visual analysis alone is not capable of doing. In the second HLM model, we introduced a predictor variable (response class) in order to help explain some of the variability in treatment responses found in the first model. Adding response class as a predictor helps us to explain about 27% of the variability in treatment effects. Though this is a significant amount of variability, there is still room to add additional predictors (e.g., age, IQ) in the future.

As hypothesized, HLM detected the greatest number of within-subject treatment effects with 33 of the 39 behaviours. In total, PEM found a treatment effect with 28 of the 39 behaviours. HLM was often more sensitive in detecting even the smallest treatment effect which accounted for this slight difference. Upon visually examining these AB graphs, it appears that in the cases where HLM and PEM did not agree, it was due to one of two factors. The first being that HLM may be so sensitive that it detects effects that while statistically significant, were not clinically significant. The second reason was that there were too few data points.

The response class there HLM and PEM both agreed that the treatment was less effective was in the arranging and ordering response class. In the arranging and ordering response class, PEM found the CBT program ineffective for 6 of the 11 behaviours, and HLM found the treatment was ineffective for 5 of 11. This result may have to do to parent follow through. The exposures for the arranging and ordering behaviours were typically not as naturally occurring as the contamination and same seat behaviours where situations such as sitting in the car happened daily. Instead arranging and ordering behaviours required more manipulation on the parents' part, such as going into their child's bedroom and specifically moving their LEGO arrangement.

Perhaps behaviours such as these with higher response effort from the parents are better treated in isolation. Future research in this area is needed.

Last, Pearson correlations were used to characterize the relative agreement across the three methods. The strongest correlation was surprisingly between PND and HLM (r = 0.69), though much different from PEM and HLM (r = 0.65). Upon reviewing the individual scores, the reason for this may be because HLM is able to detect very small treatment effects, so sometimes the HLM value is significant but still small and therefore slightly more correlated with PND. Lastly, the weakest correlation was between PEM and PND (r = 0.42). Therefore, the two comparisons involving HLM had large correlations, and the correlation between PEM and PND had a moderate correlation.

Limitations

There were several limitations to the present study that must be acknowledged. First, we only analyzed whether there was a difference between the baseline-PM phase, and active treatment phase including an FBA/I, ER/P, and CT. Visual inspection indicated a very low probability of differences being found between the baseline and PM phases, but in the future, analyses should be completed between (a) the baseline phase and the PM phase, (b) PM and the active treatment phase, and (c) the baseline phase and active treatment phase. This should be done because it is possible that PM had a cumulative effect on treatment outcomes. We instead put them together and called them baseline-PM. This was done because previous research (Vause et al., 2013, in progress) found that there was not a statistically significant difference between these two phases. Another limitation was that follow-up data was not used in the present study and therefore we are unable to speak to the long term retention of treatment effects. There is also

limited generalizability of the results given that each response class only included 11 to 14 behaviors.

With respect to the analyses, HLM does not produce an effect size. This limits comparability with other studies. In regards to an overall evaluation of the program, the treatment generally showed positive outcomes, but a statistical effect size would add value to this interpretation. Hedges single-subject effect size analogue would be a good technique to further evaluate the treatment program (Hedges, Pustejovsky, & Shadish, 2012). Another limitation is that a sensitivity analysis was not completed. This may have been beneficial given that in a couple of situation it appeared that HLM was too sensitive, and detected statistical significance in behaviours that did not show a clinically significant drop as measured by visual inspection.

Implications

The current study was the first to compare PEM, PND, and HLM for analyzing single subject research. The results show promise for the use of PEM and HLM together when analyzing similar data sets and may result in future studies which consider using HLM in addition to visual analysis when analyzing single subject research. This study shows the reader three possible ways of analyzing single subject data. We are able to see that each method has strengths and limitations, and is best suited for specific data sets. PND, for example, is less accurate at determining a treatment effect when the data is variable due to the floor/ceiling effect when a baseline point drops below. The present study illustrates how using an inappropriate method, or combination of methods, can make a treatment appear less effective than it actually is.

Lastly, using HLM allows for the researcher to include predictor variables. One implication of this is researchers would be able to better determine which population (e.g., age,

disability, IQ) their treatment package is most effective for. Future research using single subject designs should use HLM, when appropriate, to further explore possible predictor variables impacting treatment outcome.

Conclusion

Results generally showed that the *I Believe in Me, Not OCB* (Vause et al., 2013a) function-based CBT treatment program was effective in treating OCBs in children with ASD Though PND is a useful technique with certain data sets, the current data set had too much variability and could not be accurately analyzed using PND. Similar to Lumpkin et al. (2002), we found that HLM was particularly useful within this study. Overall, using HLM with PEM shows promise and should continue to be used in data sets that have enough data power for HLM to be calculated. Also, this PEM/HLM combination can be useful for data sets where the baseline data has variability.

References

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC.

Bellini, S., & Akullian, J. (2007). A meta-analysis of video modeling and video self-modeling interventions for children and adolescents with autism spectrum disorders. Exceptional Children, 73, 264–287.

Browder, D. M., Spooner, F., Ahlgrim-Delzell, L., Harris, A. A., Wakeman, S. (2008). A meta-analysis on teaching mathematics to students with significant cognitive disabilities. *Exceptional Children, 74(4),* 407-432.

Browder, D. M., & Xin, Y. P. (1998). A meta-analysis and review of sight word research and its implications for teaching functional reading to individuals with moderate and severe disabilities. *Journal of Special Education, 32(3),* 130-153. doi:10.1177/002246699803200301

Centers for Disease Control and Prevention (CDC) (2006). Prevalence of Autism Spectrum Disorders—Autism and Developmental Disabilities Monitoring Network, United States. MMWR Surveillance Summaries, 58 (SS10), 1-20

Centers for Disease Control and Prevention (CDC) (2013). Changes in Prevalence of Parent-reported Autism Spectrum Disorder in School-aged U.S. Children: 2007 to 2011–2012, United States. MMWR Surveillance Summaries, 65, 1-12.

Chen, C., & Ma, H. (2007). Effects of treatment of disruptive behaviours: A quantitative synthesis of single-subject researches using the pem approach. *The Behavior Analyst Today, 8(4),* 380-396.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied Behavior Analysis* (2nd ed.). New Jersey, NY: Pearson Education Inc.

Davis, D. (2012). Using Hierarchical Linear Modeling for the Analysis of Single-Case Data Unpublished manuscript.

Davis, D. H., Gagne, P., Fredrick, L. D., Alberto, P. A., Waugh, R. E., & Haardorfer, R. (2013). Augmenting visual analysis in single-case research with hierarchical linear modeling. *Behaviour Modification, 37(1),* 62-89. doi: 10.1177/0145445512453734.

Didden,R., Korzilius, H., Van Oorsouw, W., & Sturmey, P. (2006). Behavioral treatment of challenging behaviors in individuals with mild mental retardation: Meta-analysis of single-subject research. *American Journal of Mental Retardation, 111(4),* 290-298.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). Los Angeles, CA: Sage Publications Inc.

Feldman, M. A., Ducharme, J. M., & Case, L. (1999). Using self-instruction pictorial manuals to teach child-care skills to mothers with intellectual disabilities. *Behavior Modification, 23*, 480-497.

Fombonne, E., Zakarian, R., Bennett, A., Meng, L., & McLean-Heywood, D. (2006). Pervasive developmental disorders in Montreal, Quebec, Canada: Prevalence and links with immunizations. *Pediatrics, 118,*139–150.

Gao, Y. J. & Ma, H. H. (2006). Effectiveness of interventions influencing academic behaviors: A quantitative synthesis of single-subject researches using the pem approach. *Behavior Analysts Today. 4*, 572-422.

Hedges, L., Pustejovsky, J., Shadish, W. (2012) A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3,* 224-239.

Kazdin, A. E. (2011). Single-case research designs: Methods for clinical and applied settings (2nd ed.). New York, NY, US: Oxford University Press.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research, *65(1),* 73-93.

Lee, S.-H., Simpson, R. L., & Shogren, K. A. (2007). Effects and implications of self-management for students with autism: A meta-analysis. Focus on Autism and Other Developmental Disabilities, 22, 2–13. doi:10.1177/10883576070220010101.

Lehmkuhl, H., Storch, E., Bodfish, J., & Geffken, G. (2008). Brief report: Exposure and response prevention for obsessive compulsive disorder in a 12-year-old with autism. *Journal of Autism and Developmental Disorders, 38,* 977-981.

Levendoski, L. S.,&Cartledge,G. (2000). Self-monitoring for elementary school children with serious emotional disturbances: Classroom applications for increased academic responding. *Behavioral Disorders, 25,* 211-224.

Leyfer, O., Folstein, S., Bacalman, S., Davis, N., Dinh, E., Morgan, J., ... Lainhart, J.E. (2006). Comorbid psychiatric disorders in children with autism: interview development and rates of disorders. *Journal of Autism and Developmental Disorders, 36*(7), 849-861.

Lumpkin, P. W., Silverman, W. K., Weems, C. F., Markham, M. R., & Kurtines, W. M. (2002). Treating a heterogeneous set of anxiety disorders in youths with group cognitive behavioral therapy: A partially nonconcurrent multiple-baseline evaluation. *Behavior Therapy, 33(1),* 163-177.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification, 30* (5), *598-617. doi*: 10.1177/0145445504272974.

Ma, H. H. (2009). The effectiveness of intervention on the behavior of individuals with autism: A meta-analysis using percentage of data points exceeding the median of baseline phase (pem). *Behavior Modification, 33,* 339-359. doi: 10.1177/0145445509333173.

Mathur, S. R., Kavale, K. A., Quinn, M. M., Forness, S. R., & Rutherford, R. B. (1998). Social skills interventions with students with emotional and behavioral problems: A quantitative synthesis of single subject research. *Behavioral Disorders, 23,* 193–201.

Mirenda, P., Smith, I.M., Vaillancourt, T., Georgiades, S., Duku, E., Szatmari, P., ... Zwaigenbaum, L. (2010). Validating the repetitive behavior scale- revised in young children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 40,* 1521-1530. doi: 10.1007/s10803-010-1012-0.

O'Reilly, M. F., Green, G., & Braunling-McMorrow, D. (1990). Self-administered written prompts to teach home accident prevention skills to adults with brain injuries. *Journal of Applied Behavior Analysis*, *23*, 431-446.

Parker, R. I., & Brossart, D. F. (2006). Phase contrasts for multiphase single case intervention design. *School Psychology Quarterly, 21*, 46-61.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. The Journal of Special Education, 40, 194–204. doi:10.1177/00224669070400040101.

Preston, D., & Carter, M. (2009). A review of the efficacy of the picture exchange communication system intervention. *Journal of Autism and Developmental Disorders, 39,* 1471-1486. doi: 10.1007/s10803-009-0763-y.

Reaven, J. A., & Hepburn, S. (2003). Cognitive-behavioral treatment of obsessive-compulsive disorder in a child with Asperger syndrome: A case report. *Autism, 7,* 145.

Rodriguez, N.M., Thompson, R. H., & Stocca, C. S. (2012). Functional analysis and treatment of arranging and ordering by individuals with an autism spectrum disorder. *Journal of Applied Behavior Analysis, 45,* 1-22.

Scruggs, E. T., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22,* 221–242.

Sigafoos, J., Green, V. A., Payne, D., OReilly, M. F., & Lancioni, G. E. (2009). A classroom-based antecedent intervention reduces obsessive-repetitive behavior in an adolescent with autism. *Clinical Case Studies, 8(1),* 3-13.

Singer, J. D., & Willett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. Oxford, England: Oxford University Press.

Storch, E. A., Elysse, E. A., Lewin, A. B., Nadeau, J. M., Jones, A.M., Alessandro, S. D., ... Murphy, T. K. (2013). The effect of cognitive-behavioral therapy versus treatment as usual for anxiety in children with autism spectrum disorder: A randomized, controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry, 52(2),* 132–144.

Sze, K., & Wood, J. (2007). Cognitive behavioral treatment of comorbid anxiety disorders and social difficulties in children with high-functioning autism: A case report. *Journal of Contemporary Psychotherapy, 37,* 133-143.

Vause, T., Neil, N., Yates, H., & Feldman, M. (2013a). *I Believe in Me not OCB.* Manuscript in preparation.

Vause, T., Neil, N., Yates, H., & Feldman, M. (2013b). *I Believe in Me not OCB: Workbook.* Manuscript in preparation.

Vause, T., Neil, N., Yates, H., Jackiewicz, G., & Feldman, M. (2013). *Function-based cognitive-behavior therapy for obsessive compulsive behavior in children with autism spectrum disorder: A preliminary randomized comparison.* Manuscript in preparation.

Wood, J., Drahota, A., Sze, K., Har, K., Chiu, A., & Langer, D. (2009). Cognitive behavioral therapy for anxiety in children with autism spectrum disorders: a randomized, controlled trial. *Journal of Child Psychology and Psychiatry, 50,* 224–234.